

Tema 1. Análisis Estadístico de Datos.

Presentación y Objetivos.

La **Estadística Descriptiva** engloba una serie de técnicas de estructuración y de representación gráfica que permiten ordenar y presentar adecuadamente la información contenida en un conjunto de datos. La extrapolación de esta información para convertirla en regla aplicable a todos los datos que puedan obtenerse en circunstancias similares es el objetivo de la **Inferencia Estadística**. Entre las dos se sitúa el **Cálculo de Probabilidades** como lenguaje formal que permitirá tal extrapolación.

Los **Objetivos** de este Tema son:

1. Saber distinguir los distintos tipos de variables y datos según la escala de medida, naturaleza y representatividad.
2. Saber construir tablas de frecuencias univariantes y bivariantes.
3. Construir distribuciones marginales y condicionadas a partir de una distribución conjunta de frecuencias.
4. Conocer el concepto de independencia de dos variables.
5. Saber interpretar la información contenida en distintos tipos de representaciones gráficas.
6. Conocer qué se entiende por análisis exploratorio de datos.

Esquema Inicial.

1. Introducción.
2. Variables y datos. Tipos de datos.
3. Descripción de datos mediante tablas.
4. Descripción de datos mediante gráficos.
5. Introducción al análisis exploratorio de datos.

Desarrollo del Tema

1. Introducción

La Estadística Descriptiva comprende una serie de métodos y técnicas para:

1. Recoger y organizar datos referidos a las sucesivas observaciones de ciertos caracteres de una Población descrita previamente.
2. Esquematizar el comportamiento de las poblaciones con relación a determinados caracteres mediante tablas, gráficos o dibujos.

3. Resumir la información obtenida de las sucesivas observaciones en unos pocos datos representativos.
4. Analizar la relación de dependencia entre caracteres de una misma población.
5. Saber interpretar la información contenida en distintos tipos de representaciones gráficas.
6. Conocer alguna técnica de visualización para datos multivariantes.

2. Variables y Datos. Tipos de datos.

En Estadística, la materia prima son los datos y el producto final es el conjunto de conclusiones sobre el fenómeno de interés. Una **variable** es la característica de interés que se quiere estudiar y que toma valores diferentes en cada individuo. En general, las variables se representarán por las letras mayúsculas X, Y, Z, etc. Una variable puede tomar valores diferentes incluso en un mismo individuo si se cambian las condiciones en las que se toma la medida. Un **dato** es el valor observado de una variable en un momento dado en un individuo. Típicamente, un dato será un número (o una etiqueta en el caso de variables nominales) con un contexto, siendo ese contexto el que hace que ese número sea informativo. Por ejemplo, una variable puede ser la temperatura en un aula. Un dato sería la temperatura en el aula en este momento. Una variable podría ser las notas de la asignatura de Estadística durante este curso. Un dato sería la nota particular de Luis García. Una variable sería el tiempo que uno tarda en conectarse a Internet desde que se pincha con el ratón en el icono correspondiente. Un dato sería lo que uno tarda en conectarse ahora. Una variable sería el peso. Mi peso sería un dato.

2.1. Clasificación de los datos según su representatividad.

Según su representatividad, los datos pueden conformar toda la población o ser parte de una muestra. La **población** es el conjunto de todos los individuos del que se quiere estudiar una característica. Estos individuos pueden ser objetos, personas o las repeticiones de un experimento concreto. Una **muestra** es un subconjunto representativo de la población. Supóngase que se quiere estudiar la variable $X = \text{Tiempo de conexión a Internet desde que pincho en el icono}$. Una vez que se ha definido un contexto, las características del computador y de la conexión, la población sería el conjunto infinito de los datos que se obtendría al realizar la conexión todas las veces posibles. Una muestra sería el tiempo concreto de conexión en 50 ocasiones, en las condiciones en las que he definido este experimento. Si se quiere estudiar la variable $X = \text{Altura de los estudiantes de Informática en la Comunidad de Madrid}$, la población estaría formada por las alturas de todos los estudiantes de Informática de la Comunidad de Madrid y una muestra sería la estatura de, por ejemplo, un subconjunto de 600 alumnos tomados de todas las universidades de la Comunidad en las que pueden cursarse estos estudios.

2.3. Clasificación de los datos según su naturaleza.

Por su naturaleza, los datos pueden clasificarse en **cualitativos** y **cuantitativos**.

Los **datos cuantitativos** son números que expresan cantidades. Representan, por tanto, caracteres que pueden medirse. A su vez se dividen en **continuos**, si pueden tomar cualquier valor dentro de un intervalo real y **discretos**, si sus valores forman un conjunto numerable, finito o infinito. Generalmente estos últimos se corresponden con contar el número de veces que ocurre un suceso. Por ejemplo, si se miden peso, altura, voltaje, tiempo, longitud, velocidad, etc., se obtendrán datos cuantitativos continuos. Si se miden el número de hermanos, páginas de un libro, clientes, nº de aprobados, etc., se obtendrán datos cuantitativos discretos.

Los **datos cualitativos** son meras etiquetas o códigos que representan atributos. No se refieren a características cuantificables sino a cualidades de los individuos. Por ejemplo, profesión, estado civil, marca preferida de refresco, tipo de procesador, etc.

2.4. Clasificación de los datos según la escala de medida.

Se mide una propiedad en una persona o cosa cuando se le asigna un número para representar dicha propiedad. Mediante este proceso se pasa de tener una muestra de personas o cosas a tener un conjunto de números con cierta información. En estadística se diferenciarán cuatro escalas de medida con las que obtener datos: nominal, ordinal, de intervalo y de razón.

Las medidas tomadas en una escala **Nominal** clasifican las unidades en categorías, nada más. Características como el color de pelo, sexo o nacionalidad se miden con este tipo de escala. Se podrían asignar números a las categorías pero sería irrelevante qué números se usen, no tendrían ningún significado, serían meras etiquetas. Tampoco tendría sentido realizar operaciones con estos números, solamente se podrá decir si dos individuos u observaciones pertenecen o no a la misma categoría. Por ejemplo, se mide en una escala nominal el estado civil de una persona, que podría ser: casado, soltero, viudo, divorciado. Las marcas de los coches vendidos en un determinado mes, el tipo de carburante, etc. La escala nominal mide, por tanto, caracteres cualitativos.

En las medidas tomadas en una escala **Ordinal**, el orden de los números es importante, da algo más de información. Por ejemplo, si se sabe que el resultado en la final de 4x200 metros libres de los Campeonatos Europeos de Natación fue: 1. Italia, 2. Gran Bretaña, 3. Grecia, 4. Francia, 5. Rusia, 6. Polonia y 7. Alemania, el orden es importante, ya que Italia fue la mejor y Grecia fue mejor que Rusia. Lo único con significado es el *ranking*, el orden de los resultados. No se puede decir que Grecia fue 3 veces peor que Italia o Rusia 5 veces peor que Italia, o que la diferencia de calidad entre los equipos de Italia y Grecia es la misma que la de Rusia y

Alemania. Sólo se puede decir qué equipo es mejor que otro, sin cuantificar esa relación. Se mide con una escala ordinal cuando se recogen valoraciones de satisfacción de clientes: desde el 7. Muy satisfecho hasta el 1. Nada satisfecho. También en las encuestas en las que los alumnos valoran la actuación de un profesor: desde 5. Muy de acuerdo hasta el 1. Nada de acuerdo. La escala de Mohs, que recoge la dureza de los minerales es también una escala ordinal. Esta escala va desde el 10. Diamante (más duro) hasta el 1. Talco (menos duro).

Las escalas nominal y ordinal están asociadas con caracteres **cualitativos**. Estos caracteres representan cualidades de los individuos o cosas.

La escala de **Intervalo/Razón** es la escala más usada y familiar. Las medidas se toman en una escala de la misma unidad, como la altura en centímetros, la temperatura en grados Celsius o el tiempo de reacción en segundos. Las operaciones aritméticas con este tipo de medida sí tienen sentido. Por ejemplo, un gusano que mide 4 cm., mide dos centímetros más que uno que mide 2 cm.

Existe una diferencia más sutil entre las escalas de **Intervalo** y de **Razón**. El cero en la escala de razón tiene sentido, significa ausencia de la característica. Así, la longitud se mide en una escala de razón ya que se puede decir que el primer gusano mide el doble que el segundo y 0 cm. significa ausencia de longitud. Sin embargo, la temperatura se mide en una escala de intervalo ya que 0° no significa ni frío ni calor sino que es el punto en el que el agua pasa a estado sólido.

La escala de medida depende principalmente del proceso de medida, no de la propiedad que se mide. Así, el resultado de una prueba de natación se puede medir como quién llega primero, segundo, tercero, etc. (escala ordinal) o bien cronometrar el tiempo que tardan en recorrer la distancia requerida (escala de intervalo/razón).

Se distinguirán las medidas tomadas en una escala nominal, ordinal e intervalo/razón y se hablará indistintamente de variables o de datos cuantitativos, cualitativos, nominales, ordinales o de intervalo/razón.

3. Descripción de datos mediante tablas.

3.1 Tablas de frecuencias univariantes

Se necesita tener una idea general de cómo es el conjunto de datos para poder identificar patrones que guíen posteriores análisis. Una primera forma de resumir la información es mediante una tabla que diga qué valores diferentes se han observado y cuántos datos hay de cada valor (repeticiones). Esta tabla recibe el nombre de Tabla de frecuencias o Tabla de distribución de frecuencias (Tabla 1). En ocasiones, también uno se referirá a ella como tabla

estadística. Supóngase que se dispone de un total de n observaciones divididas en k valores o modalidades diferentes. Algunas definiciones:

- Se llama frecuencia absoluta del valor observado C_i (o modalidad C_i) al número total de individuos que presentan dicho carácter. Se denota por n_i .
- Se llama frecuencia relativa del valor observado C_i al cociente $f_i = n_i/n$.

Se verifica:

$$n = \sum_{i=1}^k n_i \quad \sum_{i=1}^k f_i = 1$$

Cuando los valores o modalidades observadas pueden ordenarse de menor a mayor, se define la frecuencia acumulada (absoluta o relativa) del valor C_i a su frecuencia sumada a las frecuencias de las modalidades anteriores. Se representará por N_i la frecuencia acumulada absoluta y F_i la relativa.

$$N_i = \sum_{j=1}^i n_j \quad F_i = \sum_{j=1}^i f_j$$

Valores	Frecuencias absolutas	Frec. absolutas acumuladas	Frecuencias relativas	Frec. relativas acumuladas
C_1	n_1	$N_1 = n_1$	$f_1 = n_1/n$	$F_1 = f_1$
C_2	n_2	$N_2 = n_1 + n_2$	$f_2 = n_2/n$	$F_2 = f_1 + f_2$
...
C_i	n_i	$N_i = n_1 + \dots + n_i$	$f_i = n_i/n$	$F_i = f_1 + \dots + f_i$
...
C_k	n_k	$N_k = n$	$f_k = n_k/n$	$F_k = 1$
TOTALES	n		1	

Tabla 1: Formato general de una Tabla de Frecuencias

Ejemplo 1: Se estudia la variable $X = \text{Número de cilindros}$ de los coches de los trabajadores de una empresa. Es una variable cuantitativa medida en una escala de intervalo/razón. Se tiene una muestra de esta variable medida en 92 coches. Su tabla de frecuencias es la siguiente:

Nº cilindros	n_i	N_i	f_i	F_i
3	3	3	0,03	0,03
4	49	52	0,53	0,56
5	2	54	0,02	0,58

6	31	85	0,34	0,92
8	7	92	0,08	1
TOTALES	92		1	

Tabla 2: Tabla del ejemplo 1

Se observa que un 53% de los coches tienen 4 cilindros y un 34% tienen 6 cilindros, que son las modalidades más frecuentes. Solamente un 5% tiene un número impar de cilindros (3 o 5) siendo estas modalidades las menos frecuentes. Los coches con 8 cilindros representan un 8 % del total.

La tabla descrita anteriormente pierde su utilidad de resumir información de manera clara y concisa cuando se tienen variables cuantitativas con muchos valores diferentes. Se tendrían tablas con muchas filas con frecuencias muy pequeñas. Esto sucederá tanto con variables continuas como con variables discretas que tengan muchos valores distintos. En este caso, se construye la tabla agrupando el rango de valores en intervalos y se determina el número de datos del conjunto que se encuentra en cada uno de ellos.

- Cada intervalo se llama **clase**. La clase i vendrá representada por su extremo superior e inferior. La unión de todos los intervalos debe recubrir todos los valores de la variable y las clases no deben solaparse.

$$(L_{i-1}, L_i]$$

- Se llama amplitud de la clase i , l_i , a la diferencia entre sus extremos. Se supondrá que esta longitud es constante.

$$l_i = L_i - L_{i-1}$$

- Se denomina **marca de clase** al punto medio del intervalo y será el valor que represente a todos los datos contenidos en ese intervalo. Se representará por x'_i .

El número de clases en que se divide el rango de un conjunto de datos se representará por k . Este número suele determinarse en función del tamaño muestral n . Algunos proponen el entero más próximo a \sqrt{n} . Otra regla conocida es la regla de Sturges en la que el número de clases es el entero más próximo a $1 + 3,3 \log_{10} n$. Generalmente, se utilizarán entre 5 y 20 clases de la misma longitud.

La tabla de frecuencias para este tipo de datos divididos en clases quedaría de la forma:

Clases	Marca de Clase	Frecuencias	Frecuencias
		absolutas	relativas
$[L_0, L_1)$	x'_1	n_1	$f_1=n_1/n$
$[L_1, L_2)$	x'_2	n_2	$f_2=n_2/n$
...
$[L_{i-1}, L_i)$	x'_i	n_i	$f_i=n_i/n$
...
$[L_{k-1}, L_k]$	x'_k	n_k	$f_k=n_k/n$
TOTALES		n	1

Tabla 3: Modelo de Tabla de frecuencias para datos agrupados

Ejemplo 2: Se estudia la variable $X = \text{Nota obtenida en la convocatoria de febrero}$ en una asignatura determinada en una muestra de 100 alumnos. Los datos originales serían: 6,33; 7,17; 2; 5,33; 8,33;... hasta 100 datos. Se consideran 10 clases que surgen naturalmente al considerar este tipo de datos, formando la tabla:

Clases	Marca de Clase	Frecuencias	Frecuencias
		absolutas	relativas
$[0, 1)$	0,5	2	0,02
$[1, 2)$	1,5	5	0,05
$[2, 3)$	2,5	10	0,1
$[3, 4)$	3,5	11	0,11
$[4, 5)$	4,5	18	0,18
$[5, 6)$	5,5	25	0,25
$[6, 7)$	6,5	15	0,15
$[7, 8)$	7,5	12	0,12
$[8, 9)$	8,5	2	0,02
$[9, 10]$	9,5	0	0
TOTALES		100	1

Tabla 4: Tabla del ejemplo 2

Se ve que la clase con más observaciones es la (5,6], con 25 datos que representan un 25% del total. Un 58% de los alumnos tienen notas entre (4,7]. Han aprobado un 54% de los alumnos y ninguno ha obtenido calificación entre 9 y 10.

3.2 Tablas de frecuencias bivariantes.

Cuando de cada individuo se observan dos o más variables, se obtiene un conjunto de datos multivariantes. En concreto, con dos características por individuo se tendría un conjunto de datos bivariantes. Por ejemplo, se recogen las notas en Matemáticas y Literatura de una muestra de alumnos de Bachillerato. Se tendría entonces un conjunto de datos de la forma (3,6), (5,7)...., donde la primera nota es la de Matemáticas y la segunda la de Literatura. De esta forma, el primer alumno de la muestra ha obtenido un 3 en Matemáticas y un 6 en Literatura y así sucesivamente.

El planteamiento general es el siguiente:

Sea una muestra de tamaño n descrita por las variables X e Y , o, de forma equivalente, sea un conjunto de datos bivariantes $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$. Se designará por a_1, a_2, \dots, a_k y por b_1, b_2, \dots, b_p , los k y p valores distintos que pueden tomar X e Y respectivamente. Si alguna de estas variables fuese continua o tomara demasiados valores distintos, estos valores representarán las correspondientes marcas de clase una vez que se hayan agrupado los datos en clases.

Existen varias formas de estudiar las repeticiones en una serie de datos bivariantes o bidimensionales:

- Considerando ambas medidas de forma simultánea (**distribución conjunta**).
- Considerando cada variable X e Y por separado (**distribuciones marginales**).
- Fijando el valor de una de las variables y estudiando los valores de la otra (**distribuciones condicionadas**).

3.2.1. Distribución conjunta.

Se representará por n_{ij} el número de elementos de la muestra que presentan el valor (a_i, b_j) , es decir, la frecuencia absoluta del valor (a_i, b_j) . Si se representa esta distribución conjunta en una tabla de doble entrada, cada dimensión de la tabla se corresponderá con una variable y cada celda de la tabla tendrá el número de individuos que tengan los valores correspondientes según la fila y la columna en que se encuentren. Este valor será la frecuencia conjunta.

$X \setminus Y$	b_1	b_2	...	b_j	...	b_p
a_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1p}
a_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2p}
...
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ip}
...
a_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kp}

Tabla 5: Distribución conjunta

Esta tabla puede definirse también utilizando las frecuencias relativas. Las relaciones que se verifican en estas tablas de doble entrada son las siguientes:

$$f_{ij} = \frac{n_{ij}}{n} \quad \sum_{i=1}^k \sum_{j=1}^p n_{ij} = n \quad \sum_{i=1}^k \sum_{j=1}^p f_{ij} = 1$$

Si ambas variables son cualitativas (nominales u ordinales), la tabla X-Y recibe el nombre de **tabla de contingencia**.

Ejemplo 3: En una muestra de 90 estudiantes se recogen las variables $X = \text{Número de horas semanales de estudio de una asignatura}$ e $Y = \text{Calificación final en esa asignatura}$. La variable Y, en lugar de medirla en una escala de intervalo/razón, se va a medir en una escala ordinal con las categorías o modalidades: Suspenso (S), Aprobado (A), Notable (N) y Sobresaliente (B). El conjunto de datos original sería: (3,S), (4,N), (2,A) ..., una por cada estudiante de la muestra. Se puede disponer toda la información de la muestra en una tabla de distribución conjunta (absoluta):

X \ Y	S	A	N	B
1	19	0	0	0
2	10	16	1	0
3	6	13	4	0
4	5	3	3	1
5	0	3	2	4

Tabla 6: Tabla del ejemplo 3

3.2.2. Distribución marginal y condicionada.

El estudio de la distribución marginal de cualquiera de las variables solamente tiene sentido partiendo de las tablas de distribución conjunta descritas en el apartado anterior. A partir de ellas se quiere estudiar qué ocurre si uno se olvida de una de las variables y se centra en la otra. La tabla siguiente ilustra la distribución marginal para las dos variables X e Y.

X \ Y	b ₁	b ₂	...	b _j	...	b _p	Marginal X
a ₁	n ₁₁	n ₁₂	...	n _{1j}	...	n _{1p}	$n_{1\bullet} = \sum_{j=1}^p n_{1j}$
a ₂	n ₂₁	n ₂₂	...	n _{2j}	...	n _{2p}	$n_{2\bullet} = \sum_{j=1}^p n_{2j}$
...
a _i	n _{i1}	n _{i2}	...	n _{ij}	...	n _{ip}	$n_{i\bullet} = \sum_{j=1}^p n_{ij}$
...
a _k	n _{k1}	n _{k2}	...	n _{kj}	...	n _{kp}	$n_{k\bullet} = \sum_{j=1}^p n_{kj}$
Marginal				

Y	$n_{\bullet 1} = \sum_{i=1}^k n_{i1}$	$n_{\bullet 2} = \sum_{i=1}^k n_{i2}$	$n_{\bullet j} = \sum_{i=1}^k n_{ij}$	$n_{\bullet p} = \sum_{i=1}^k n_{ip}$	$n = \sum_{i=1}^k \sum_{j=1}^p n_{ij}$
---	---------------------------------------	---------------------------------------	---------------------------------------	---------------------------------------	--

Tabla 7: Distribución conjunta y marginales

El nombre de marginal viene de la localización de estas distribuciones en los márgenes de la tabla de frecuencias conjuntas. Se utilizará la siguiente notación:

$n_{i\bullet} = \sum_{j=1}^p n_{ij}$ representa la frecuencia absoluta asociada al resultado a_i y $f_{i\bullet} = \sum_{j=1}^p f_{ij} = \frac{n_{i\bullet}}{n}$ su

frecuencia relativa. Igualmente, $n_{\bullet j} = \sum_{i=1}^k n_{ij}$ representa la frecuencia absoluta del resultado b_j y

$f_{\bullet j} = \sum_{i=1}^k f_{ij} = \frac{n_{\bullet j}}{n}$ su frecuencia relativa.

Ejemplo 4: Se completa la tabla del ejemplo anterior añadiendo las distribuciones marginales.

X \ Y	S	A	N	B	Marginal X
1	19	0	0	0	19
2	10	16	1	0	27
3	6	13	4	0	23
4	5	3	3	1	12
5	0	3	2	4	9
Marginal Y	40	35	10	5	90

Tabla 8: Tabla del ejemplo 4

Lo que significa que si se estudia por separado la variable $X = \text{Número de horas semanales de estudio de una asignatura}$, en su distribución marginal de frecuencias se observa que, de 90 estudiantes encuestados, 19 estudiaron 1 hora/semana, 27 estudiaron 2 horas/semana y así sucesivamente. Del mismo modo, si se estudia la variable $Y = \text{Calificación final en esa asignatura}$, 40 alumnos suspendieron, 35 sacaron aprobado, etc.

Si se fija el valor de una de las variables, ¿cómo se distribuye la otra? Supóngase, por ejemplo, que en la tabla anterior se fija $X=4$, restringiendo el estudio a los alumnos que estudiaron 4 horas/semana. ¿Cuál es la distribución de la Y ahora? Se tendrían un total 12 alumnos que estudiaron 4 horas/semana, de los cuales 5 suspendieron, 3 aprobaron, 3 sacaron notable y 1 sobresaliente.

A este proceso de fijar el valor de una de las variables se le denomina condicionar y equivale a restringir el estudio descriptivo a un subconjunto de la muestra o población inicial. Así, la distribución de X condicionada, por ejemplo, por el valor $Y=b_j$, tendría la siguiente estructura:

X Y=b_j	Frecuencias absolutas	Frecuencias relativas
a₁	n_{1j}	$f_1^j = \frac{n_{1j}}{n_{\bullet j}}$
a₂	n_{2j}	$f_2^j = \frac{n_{2j}}{n_{\bullet j}}$
...	...	
a_i	n_{ij}	$f_i^j = \frac{n_{ij}}{n_{\bullet j}}$
...	...	
a_k	n_{kj}	$f_k^j = \frac{n_{kj}}{n_{\bullet j}}$
TOTALES	$n_{\bullet j}$	1

Tabla 9: Distribución de frecuencias para X|Y=b_j

Se llamará frecuencia relativa de la modalidad a_i de X condicionada a la modalidad b_j de Y a:

$$f_i^j = f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$$

Del mismo modo, la frecuencia relativa de la modalidad b_j de Y condicionada a la modalidad a_i de X será:

$$f_j^i = f_{j|i} = \frac{n_{ij}}{n_{i\bullet}}$$

Ejemplo 5: En el ejemplo de las horas de estudio y las notas, se construye la tabla de frecuencias para la distribución de Y | X=4:

Y X=4	Frecuencias absolutas	Frecuencias relativas
S	5	0,42
A	3	0,25
N	3	0,25
B	1	0,08
TOTALES	12	1

Tabla 10: Tabla del ejemplo 5

Se verifica que:

$$\text{CONJUNTA} = \text{MARGINAL} \times \text{CONDICIONADA}$$

$$f_{ij} = f_j^i f_{i\bullet} = f_i^j f_{\bullet j}$$

3.2.3. Independencia.

Se dirá que el carácter o variable X es **independiente** del carácter Y si todas las distribuciones condicionadas $X|Y=b_j$ son idénticas independientemente del valor de Y. Es decir, para cualquier i,

$$f_i^j = f_{i\bullet} \text{ para todo } j, \text{ y no es función de } j.$$

La independencia es siempre recíproca. Cuando X e Y sean independientes, se verificará que la distribución conjunta será el producto de las marginales, es decir:

$$f_{ij} = f_{i\bullet} f_{\bullet j} \text{ para todo } i, j.$$

En la tabla estadística, la independencia se traduce en:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} \text{ para todo } i, j.$$

4. Descripción de datos mediante gráficos.

Además de las tablas ya descritas, las distribuciones de frecuencia pueden estructurarse en gráficos. Sin duda es la forma más eficaz y rápida, si se dispone de las herramientas adecuadas, de resumir la información de un conjunto de datos. Hay muchas formas de realizar representaciones gráficas. En esta sección se van a considerar las más habituales.

4.1 Diagrama de barras.

Es la representación gráfica de una tabla de frecuencias en la que los datos están sin agrupar. Consiste en dibujar un rectángulo por cada valor de la variable, con área proporcional a su frecuencia. Es útil para variables cualitativas (nominales y ordinales) o cuantitativas discretas con pocos valores diferentes. El diagrama de barras de la figura 1 muestra la distribución de las ventas de turismos por marcas en España en mayo del 2006. También se muestra en la figura 2 el diagrama de barras para la tabla de frecuencias del ejemplo 1 de los cilindros de los coches.

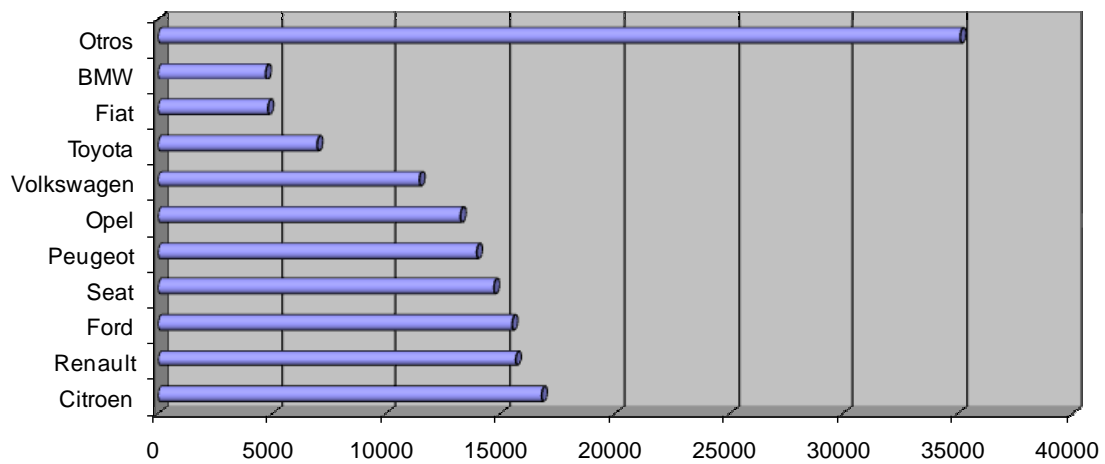


Figura 1: Diagrama de barras para las ventas de turismos

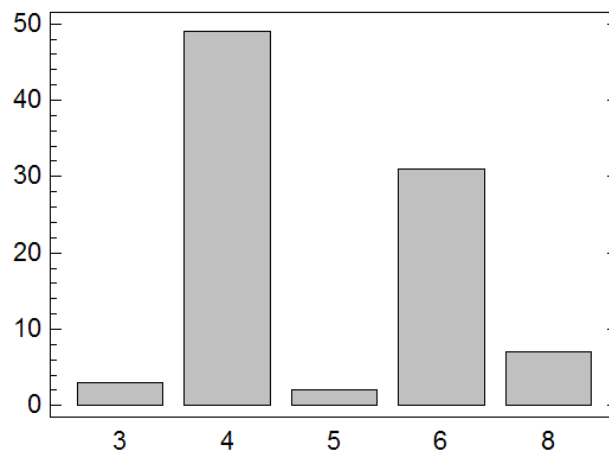


Figura 2: Diagrama de barras del ejemplo de los cilindros

4.2. Diagrama de sectores o diagrama de tarta.

Se utiliza también cuando la variable tiene pocos valores diferentes. Se construye dividiendo un círculo en sectores con áreas proporcionales a la frecuencia de cada valor, de forma que la suma del área de todos los sectores es el área del círculo. La figura 3 muestra dos diagramas de tarta o sectores.

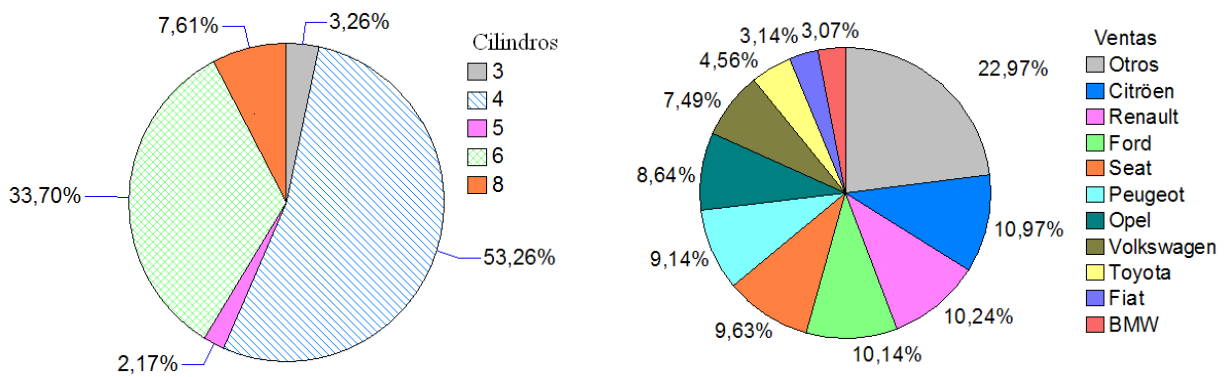


Figura 3: Diagrama de tarta para los datos de cilindros y de ventas de coches

4.3. Histograma y Polígono de Frecuencias.

Un Histograma es la representación gráfica de una tabla de frecuencias en las que los datos han sido agrupados en intervalos o clases. Se utiliza pues para variables cuantitativas que toman muchos valores diferentes. Cada rectángulo corresponde a una clase y su área es proporcional a la frecuencia de dicha clase.

En un histograma uno se debe fijar en diversos aspectos:

- **Concentraciones:** aquéllos rectángulos de mayor altura, en los que hay mayor proporción de datos y en torno a los que se disponen otros de frecuencia inferior o decreciente.
- **Huecos:** podrían ser un indicio de que se mezclan datos de poblaciones diferentes.
- **Valores atípicos:** en el tema siguiente se verá una regla para detectarlos. En general, un dato será atípico si se separa del patrón general de los datos, tanto si es muy grande como muy pequeño.
- **Asimetrías:** indican hacia dónde tienden a desplazarse los datos cuando uno se aleja de las zonas de concentración. Por ejemplo, cuando la cola de la distribución de los datos apunta hacia la derecha, se dice que la asimetría es positiva.

Es aconsejable hacer varios histogramas cambiando el número de clases para comprobar que las características que se observan no se deben a un agrupamiento casual de los datos.

Ejemplo 6: En la figura 4 se muestra el histograma correspondiente a la variable $X = \text{Precio}$, medida en la muestra de los 92 coches del ejemplo 1. Se han dividido los valores para el precio en intervalos que van desde el 5 hasta el 65 (en miles de euros). En total 10 intervalos de longitud 6. Se observa que la clase más frecuente es la comprendida entre 11.000 y 17.000 euros y que a partir de ahí las frecuencias van disminuyendo conforme aumenta el precio. Las clases entre 41.000 y 47.000 euros y entre 53.000 y 59.000 euros carecen de observaciones, lo que

podría ser indicio de que se están mezclando datos de dos poblaciones diferentes. Los datos más alejados, los que se encuentran en las clases entre 47.000 y 53.000 euros y 59.000 y 65.000 euros, no deben considerarse como atípicos ya que no se salen del patrón general de la distribución que es una asimetría hacia valores altos. La distribución presenta asimetría positiva.

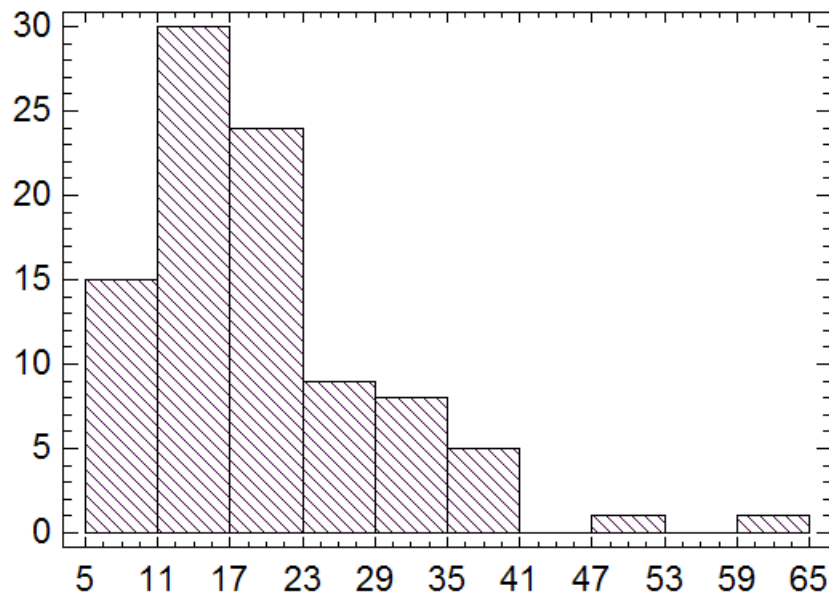


Figura 4: Histograma de precios de coches

El **polígono de frecuencias** es la línea poligonal que resulta de unir los puntos medios de la parte superior de los rectángulos en el histograma. En ocasiones, sobre todo con tamaños muestrales grandes, el polígono de frecuencias puede ayudar a hacerse una idea más clara de cómo son los datos. La figura 5 muestra el polígono de frecuencias para los datos de los precios de los coches.

El **polígono de frecuencias acumuladas** se define a partir de la tabla de frecuencias para una variable continua cuyos valores han sido agrupados en clases. Es la línea que resulta de unir los pares de valores $(L_{i+1}, F(L_{i+1}))$, es decir, los extremos superiores de las clases y la frecuencia acumulada hasta ese valor.

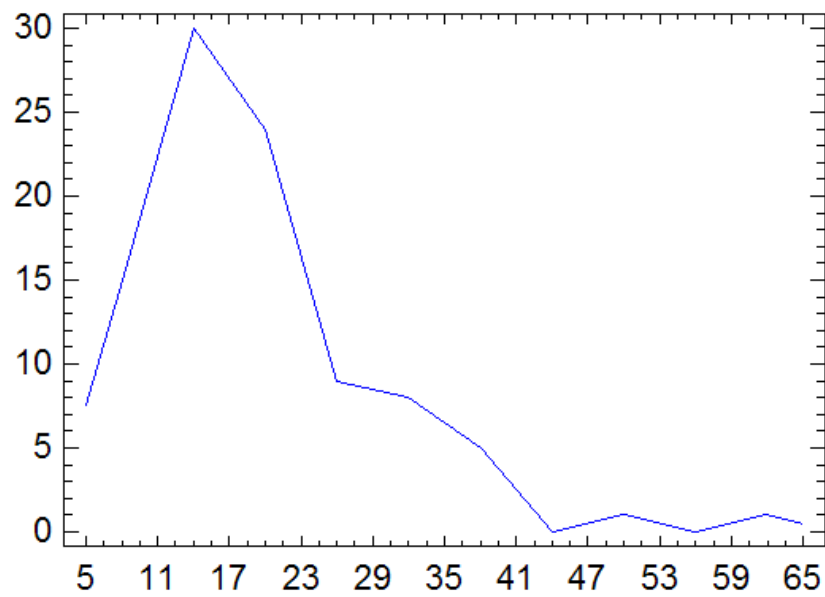


Figura 5: Polígono de frecuencias para los precios de coches.

A partir de la tabla 4 de frecuencias del ejemplo 2 se tiene el polígono de frecuencias acumuladas de la Figura 6. En él se han representado los pares de datos (1; 0,02), (2; 0,07), (3; 0,17), (4; 0,28) etc.

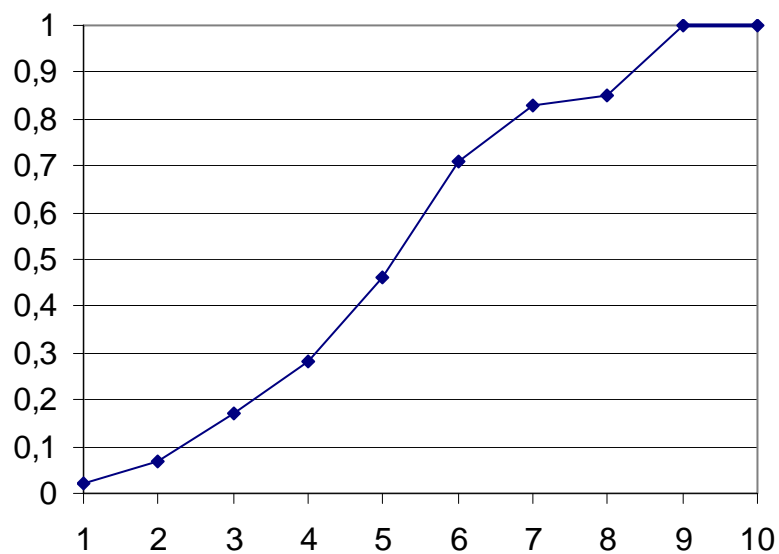


Figura 6: Polígono de frecuencias acumuladas

4.4. Diagrama de tallo-hojas.

Este tipo de diagrama fue descrito por Tukey y es utilizado para representar distribuciones de variables cuantitativas. Además, en la misma representación gráfica se visualizan los valores que se estudian. Los pasos para construirlo se ilustran con un ejemplo.

Sea una muestra de la variable $X = \text{Peso en Kg}$ en un grupo de 60 estudiantes: 54, 60, 62, 53, etc.

- Se redondean los datos a dos o tres cifras, expresando los valores con números enteros. Pueden expresarse en otras unidades (dividiendo o multiplicando) para que al redondear puedan obtenerse cifras de decenas o centenas repetidas. En el ejemplo, como se tienen datos de dos cifras, se dejan como están.
- Se ordenan los datos de menor a mayor.

44, 45, 46, 46, 47, 48, 49, 50, 50, 50, 52, 52, 52, 52, 53, 53, 53, 54, 54, 54, 55, 55, 55, 55, 56, 56, 56, 57, 60, 60, 60, 60, 60, 61, 61, 62, 62, 63, 64, 64, 64, 65, 65, 65, 66, 67, 68, 68, 68, 70, 70, 70, 70, 71, 72, 72, 74, 75, 80, 93.

- Se separan por la izquierda uno o más dígitos de cada dato, según sea el número de filas que se quiera obtener, normalmente no más de 12 o 15. Cada uno de estos valores se escribe uno debajo del otro, trazando una línea a la derecha de los números escritos. Estas cifras constituyen el tallo. En el ejemplo, se tomará la primera cifra.
- Para cada dato original se busca el dígito del tronco y a la derecha de la línea se escriben las cifras que habían quedado, cifras que forman las hojas.

Se obtendría el gráfico:

4		4	5	6	6	7	8	9														
5		0	0	0	2	2	2	2	3	3	3	4	4	4	5	5	5	5	6	6	6	7
6		0	0	0	0	0	1	1	2	2	3	4	4	4	5	5	5	6	7	8	8	8
7		0	0	0	0	1	2	2	4	5												
8		0																				
9		3																				

Figura 7: Gráfico de tallo-hojas

El resultado es, básicamente, un histograma *tumbado* con longitud de las clases igual a 10 en el que, además de mostrarse la forma de la distribución, se pueden visualizar los datos.

Ejemplo 7: Sea la variable $X = \text{Peso}$ recogida en la muestra de coches del ejercicio 1. La figura 7 muestra el diagrama de tallo y hojas que se obtiene con un software estadístico convencional. Los datos originales eran, en libras: 1695, 1845, 1965, 2045, 2055, 2240, 2270, 2285, 2295, 2297, etc. En la representación se observa que se han redondeado los números perdiendo las

cifras de centenas y unidades. Así, el dato 1695 se redondea a 1600 y se representan la primera cifra en el primer tronco como 1| y la hoja como 6. Los números que figuran en la columna de la izquierda representan las frecuencias absolutas acumuladas hasta la clase que contiene la mediana que se señala con un paréntesis. Este tipo de diagrama es muy útil para el cálculo de algunas medidas características que se estudiarán en el tema siguiente.

Gráfico de tallo-hojas para Peso: unidad = 100,0 1|2 representa 1200,0

1	1 6
3	1 89
5	2 00
14	2 222223333
23	2 444444555
31	2 66667777
45	2 88888889999999
(7)	3 0000001
41	3 2222333
34	3 4444444555555
21	3 66667777777
10	3 89999
5	4 00011

Figura 8: Gráfico de tallo-hojas

5. Introducción al Análisis Exploratorio de Datos (AED)

Cuando todas las técnicas de tabulación y representación gráfica que se han visto se utilizan no solamente con el propósito de describir un conjunto de datos sino como un medio para descubrir la información oculta en los mismos, se inicia el **análisis exploratorio de datos**, introducido por Tukey en 1977. No es una técnica paralela a las que se han visto sino una aproximación o filosofía para el análisis de datos que emplea una variedad de técnicas para:

- Profundizar lo más posible en el conocimiento de un conjunto de datos
- Descubrir estructuras y relaciones entre las variables
- Detectar variables de interés en el estudio
- Detectar valores anómalos o atípicos
- Comprobar hipótesis acerca de los datos
- Diseñar modelos que describan los datos

En lugar de contrastar en un conjunto de datos una serie de hipótesis clásicas, predeterminadas de antemano, el AED dice cómo se tienen que diseccionar los datos para que ellos mismos revelen su estructura, patrones y comportamiento. Es decir, cómo buscar, qué buscar y cómo interpretar lo encontrado.

Todo lo visto en este tema son técnicas utilizadas en AED. Véase otro tipo de gráfico que se utiliza cuando se quieren estudiar dos o más variables medidas sobre el mismo individuo, para hacerse una idea de qué tipo de relación existe entre ellas, si existe alguna. Se estudiarán más técnicas del AED en próximos temas.

5.1. Diagrama de dispersión.

Ayuda a ver la relación que puede existir entre dos variables X e Y . Es simplemente una gráfica en la que en el eje horizontal se representan los valores de la primera variable y en el eje vertical los valores de la segunda. Se tendrán tantos puntos como tamaño de la muestra.

Ejemplo 8: En una muestra de 130 personas se recogen los valores de las variables $X = \text{Temperatura (en } ^\circ \text{Fahrenheit)}$ e $Y = \text{Pulsaciones por minuto}$. Se representan estos 130 pares de datos en un diagrama de dispersión en la figura 8. Aparentemente, la nube de puntos que resulta no permite ver ningún tipo de relación que destaque entre estas dos variables. Sí se puede ver que parece que un dato se sitúa más a la derecha que el resto, se podría estudiar para ver si es un dato atípico.

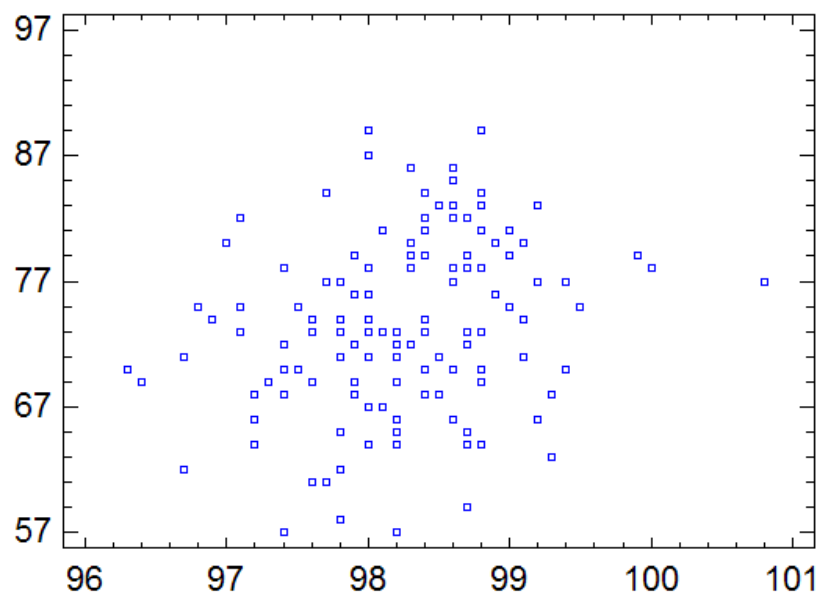


Figura 9: Diagrama de dispersión de Temperatura frente a Pulsaciones

Ejemplo 9: Se tiene en un fichero información referente a las variables $X = \text{Millas por galón de gasolina en ciudad}$ (el equivalente americano al Km. por litro de gasolina europeo), $Y = \text{Millas por galón en autopista}$ y $Z = \text{Potencia}$ en la muestra de coches del ejemplo 1. Se puede hacer una matriz de diagramas de dispersión que representará este diagrama para todos los pares de variables del fichero. Este tipo de matriz puede verse en la figura 9. Se observa rápidamente que los valores de las variables X e Y están prácticamente dispuestos a lo largo de

una línea recta, sugiriendo un tipo de relación lineal entre ambas variables. Además, un valor alto en X implica también un valor alto en Y (gráfico 1). Por otra parte, la relación entre las variables Z e Y no parece ser lineal a juzgar por el perfil curvo que parecen dibujar los datos (gráfico 2).

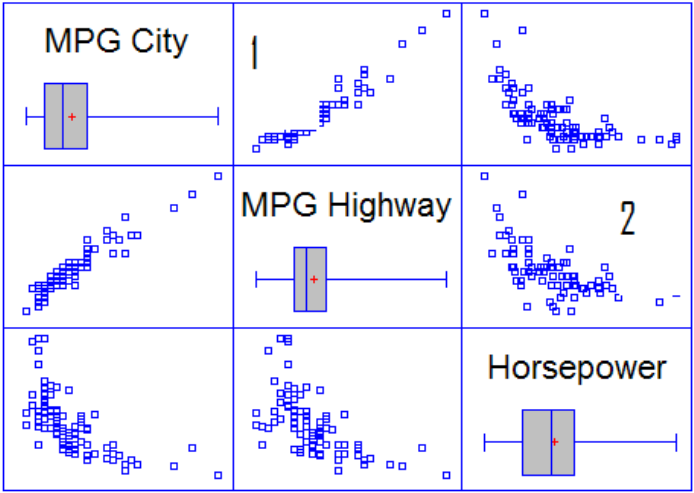


Figura 10: Matriz de diagramas de dispersión